DOCUMENT RESUME

ED 157 434 HE 010 256

AUTHOR Halpern, Sheldon; And Others
TITLE Evaluation of Student Writing: A Comparative Study.
INSTITUTION Bowling Green State Univ., Ohio.
PUB DATE Jun 77
NOTE 11p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Basic Skills; *College Faculty; *College Freshmen; Comparative Analysis; English Departments; Evaluation Criteria; Higher Education; Institutional Research; Research Projects; Student Evaluation; *Writing Skills
IDENTIFIERS *Bowling Green State University OH

ABSTRACT

A study of the reliability of writing evaluation by teachers from various departments at Bowling Green State University is reported. One hundred papers from the Writing Proficiency Examination given by the Department of English to all students completing the freshman composition sequence were selected for evaluation. Thirty volunteer faculty members were used to make up 10 three-member groups. Nine of the 10 groups comprised faculty from three different areas (Business, Education, Humanities, Science, and Social Science); the tenth group comprised one faculty member from education and two from Social Science (Psychology and Geography). Results of the evaluation and a comparison with the assessment made by the English Department are provided. (SPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Bowling Green State University

Department of English
Bowling Green, Ohio 43403

ED157434

EVALUATION OF STUDENT WRITING: A COMPARATIVE STUDY

Sheldon Halpern
Elmer Spreitzer
Stuart Givens

A Report To

Academic Council, Bowling Green State University

June 1, 1977

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Elmer A. Spreitzer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

HE010256

The problems of student literacy, or the lack of it, have been addressed by almost every college, university, and school system in the nation with a variety of programs involving English teachers at one end and testers at the other. The goals of in-service and pre-service training, development of methodologies, tutorial clinics, and reinstatement of the traditional year-long freshman composition course are to provide students with clearly defined skills and competencies in handling information and writing articulate standard English prose. At the same time, performance on one or another test of language proficiency is being established as a criterion for high school graduation, college admission, degree completion, and other decision points in the student's academic career.

One factor, almost always identified as contributing to the problem of student language skill but rarely dealt with in proposed solutions, is the lack of reinforcement of reading and writing competency in student experiences outside the English classroom. Although it is still a tenet of academia that writing is the ultimate tool of scholarly achievement, there is far less demand on students to write than most academicians think appropriate, and the writing that is done is usually not evaluated for the quality of composition. The reason often cited for this obvious discrepancy between theory and practice is that teachers untrained in language disciplines are unqualified to judge student writing.

There are at least three distinct steps in improving student writing; evaluation of the student's competency, diagnosis of the problems, and remediation of the weaknesses.

We believe that most, if not all, college faculty are qualified to do the first and many are also competent to do the second, even if they are correct in leaving

most diagnosis and all remediation to their colleagues in English and communication areas. For example, an Associate Professor of Sociology remarked, "When I get a badly written paper, I don't know how to tell the student what kinds of sentence structure errors he made or how to correct them. What I do know is that he's using the right textbook words, but I can't tell from the paper whether or not he knows what he is talking about." Clearly this instructor evaluated the writing as bad and diagnosed at least one problem as being poor sentence structure. More importantly he identified the student's weakness in writing as being inseparable from the adequacy of his performance in Sociology. A few written comments to this effect, especially if they are reflected in grades, should be sufficient to send the student to the Writing Clinic to seek help; or at least to motivate him to polish some composition skills that had been left to rust since completion of freshman English.

Yet such comments are relatively rare. In the Spring of 1976, as part of a general study of writing proficiency, we surveyed a number of juniors and seniors at Bowling Green State University about their writing experiences in courses outside freshman composition. They reported having been required to write either essay examinations or course papers in only 43% of their classes. Although at least some of the papers were returned in 87.5% of these courses, evaluative comments were written on them only 57% of the time. This means that three-quarters of all instruction was provided with no indication to the students that their ability to express themselves in writing was of importance!

In order to determine whether such reluctance on the part of the faculty was legitimate, we did a study of the reliability of writing evaluation by teachers from various departments. As a control, we utilized the Writing Proficiency Examination given by the Department of English to all students completing the freshman composition sequence. In order to meet a University-wide undergraduate

degree requirement, students must pass this program-wide uniform theme-writing examination. It is given at the end of each quarter, and students are allowed to take it only if they are doing minimally passing work in their composition classes. Papers are each graded twice in randomly selected batches by composition instructors against a five-dimensional evaluation rubric with fixed standards. In order to pass, a paper must receive passing grades in all five dimensions from two graders. When the two grades disagree, a third reading by a highly experienced instructor breaks the tie.

Experimental graders were chosen by inviting fifty randomly selected full-time teaching faculty to participate. Thirty-two volunteered, and thirty were used to make up ten three-member groups. Nine of the ten groups comprised faculty from three different areas (Business, Education, Humanities, Science, and Social Science); the tenth group comprised one faculty member from Education and two from Social Science (Psychology and Geography). Experimental graders were not informed of one another's identity.

A random sample of one hundred papers was selected from all proficiency examinations written at the end of Winter Quarter, 1977, stratified by the following categories:

1. Two grades of Pass
2. Two grades of Pass, one Fail
3. Two grades of Fail, one Pass
4. Two grades of Fail

In order to represent the proportions in the entire population of Winter, 1977 examinations, the sample contained sixty-nine papers of category one, eleven of category two, and ten each of categories three and four. Ten batches of ten papers each were made up, similarly distributed among categories.

Each batch of papers was then given in successive weeks to each of the three members of the experimental groups, with the following instructions:

Here are the ten papers you are to grade. PLEASE DO NOT PUT ANY MARKS ON THE PAPERS THEMSELVES, WHICH WILL BE GRADED BY OTHER PARTICIPANTS.

Please note that each paper has been assigned a code number and has attached to it the topic on which it was written. In evaluating the paper, use whatever criteria seem appropriate to you to apply to Bowling Green undergraduates. Try to make your standards reasonably rigorous, aiming at a level of proficiency average students should realistically meet after a general studies program in English composition. Please accept the assigned topic as an appropriate subject for writing.

Please read through all ten papers once, and then reread each and score it on the attached score sheet as follows:

Column 1 - Score from 1 (low) to 9 (high), with 5 as the marginal point, for Handling of Information including such considerations as organization, logic, specificity, and relevance to the assigned topic.

Column 2 - Score from 1 to 9 on the same scale, for Handling of Language, including such considerations as sentence structure, vocabulary, clarity, and mechanics of syntax, punctuation, and spelling.

Column 3 - Score S if the paper meets your minimal standards for acceptable undergraduate writing, and U if it does not.

Column 4 - Enter any comments you care to on the paper's strengths or weaknesses.

The resulting S/U scores were used to place each paper in one of the four following experimental categories, parallel to the English categories above:

1. SSS
2. SSU
3. SUU
4. UUU

Table I shows the extent of congruence between the two sets of evaluations.

Congruence →	Experimental Category 1	Experimental Category 2	Experimental Category 3	Experimental Category 4	
English Category 1	41	23	4	1	PASS
English Category 2	5	4	1	1	
English Category 3	2	5	3	0	FAIL
English Category 4	2	3	4	1	
	PASS		FAIL		

As can be seen, 49% of the papers were placed in the same categories by both sets of readers, whereas random placement would have produced only 25%. More strikingly, if we consider Table I on a dichotomous Pass/Fail basis, there is 81% agreement between the two sets of graders in applying their standards for minimal acceptability. It also seems reasonable to include within the area of general agreement the six papers which were placed in Category 2 by one group and Category 3 by the other, since there is not a great deal of judgmental difference between a 2 to 1 opinion on one side and a 1 to 2 opinion on the other. Also worth noting is the fact that only three papers fall into the extreme positions of being unanimously passed by one group and unanimously failed by the other.

Because the assessment by the English Department determines whether or not the student meets the proficiency standards of the University, it can be viewed as the "correct" judgment in each paper. In that case, the general faculty succeeded in identifying competency in 73 of 80 cases (91.25%) and in identifying incompetency in 8 of 20 cases (40%). This discrepancy can, we believe, be largely explained by further analysis of the standards of evaluation used by the experimental graders.

Table II shows the mean scores assigned by the experimental graders for all papers in each of the four quadrants of Table I.

	Experimental Pass	Experimental Fail
English Pass	n = 73 Information $\bar{X}$ = 6.6 Language $\bar{X}$ = 6.4	n = 7 Information $\bar{X}$ = 5.3 Language $\bar{X}$ = 5.0
English Fail	n = 12 Information $\bar{X}$ = 6.1 Language $\bar{X}$ = 5.3	n = 8 Information $\bar{X}$ = 4.7 Language $\bar{X}$ = 4.1

Referring back to the scoring instructions given to the experimental graders, it can be inferred that they considered papers as unsatisfactory only when they perceived inadequacy in both handling of language and handling of information. The twelve papers in the lower left quadrant were not scored much higher on language than the seven in the upper right, but were scored considerably higher on information. It is reasonable to infer, then, that the general faculty did not necessarily arrive at "incorrect" judgments on these papers, but simply was more tolerant than the English faculty of perceived language weaknesses in papers which were relatively strong in content. This inference is supported by gamma coefficient analysis of correlation between the information and language scores given each paper by the three experimental graders (one to nine ascending) and that paper's placement in an experimental category (one to four descending). The gammas were .58 for the Handling of Information score and .56 for Handling of Language, the small difference indicating a slight weighting in favor of content. Such standards on the part of the general faculty are both expected and appropriate; frankly, we had expected even a heavier weighting in this direction.

The next stage of analysis was to determine what relationship, if any, existed among three variables: the departmental affiliation of the experimental grader, the extent of agreement between the experimental grader's judgments and those of the English Department, and the quantity and types of comments made on the papers. Table III lists the experimental graders in four groups according to extent of agreement. The Agreement Group comprises those twelve graders who assigned

grades of S to seven or eight of the eight papers passed by English and grades of U to one or both papers failed by English. The Disagreement Group comprises those five graders whose grades disagreed with those of English on three or more papers, but who also assigned between seven and nine grades of S; since each batch contained eight papers passed by English, the members of this group agreed substantially with the proportion of passing papers though not with the identification of those papers. The More Stringent group comprises those seven graders who assigned more than three grades of U and ranges from four to eight such grades. The Less Stringent group comprises those six graders who assigned ten grades of S.

Within each Group, graders are listed in descending order of number of comments. The Column marked "Number" indicates the number of papers (out of ten) about which the grader made any discursive comments at all. The next two columns show the number of papers about which the grader's comments indicated an evaluation of Information Handling or Language Handling, respectively. The last column accounts for comments not assignable to either of the two previous columns (e.g., "Poor handwriting" or "Marginal performance"). Please note that the rows do not add up any specific number, since comments on any paper may deal with Information, Language, both, or neither.

Table III

Grouping		Department	Number	Information	Language	Not Assignable
Agreement	1	Chemistry	10	1	10	0
	2	Business Education	10	10	4	0
	3	Psychology	7	3	4	2
	4	Geography	6	3	3	2
	5	Home Economics	5	2	0	3
	6	Geology	3	0	0	3
	7	Business Education	2	0	0	2
	8	Business Education	0			
	9	Library & Educ. Media	0			
	10	Romance Languages	0			
	11	Sociology	0			
	12	Geology	0			
		Sub-totals	43	19	21	12
Disagreement	13	Accounting	10	10	9	0
	14	Physical Education	10	7	10	0
	15	History	8	3	3	2
	16	English	3	0	0	3
	17	Economics	3	2	1	0
		Sub-totals	34	22	23	5
More Stringent	18	English	10	7	4	2
	19	Psychology	10	5	9	1
	20	Romance Language	9	4	3	3
	21	History	8	7	2	1
	22	Home Economics	3	0	2	1
	23	Speech (Theatre)	1	0	0	1
	24	Economics	1	0	0	1
		Sub-totals	42	23	20	10
Less Stringent	25	Political Science	10	6	9	0
	26	History	9	1	1	7
	27	Geography	7	3	3	1
	28	Industrial Education	2	1	0	1
	29	History	2	2	0	0
	30	Special Education	0			
		Sub-totals	30	13	13	9
		Totals	149	77	77	36

Table III shows a fairly even distribution of faculty evaluative effort. Given a total of three hundred readings, the faculty commented on information and language one-fourth of the time each, and in more general ways one-eighth of the time. Half of the papers were evaluated with no comment at all. The group which agreed most consistently with the judgments of the Department of English actually commented least, averaging 3.58 papers each. Furthermore, faculty members from empirical or professional areas were as likely to make comments on both form and content, as likely to agree or disagree with the judgments of the English Department, and as likely to have more or less stringent standards as their colleagues in the humanities.

The inference to be drawn is clear. Independent of their disciplines or their sense of confidence in diagnosing and correcting writing problems, faculty are perfectly capable of making valid judgments about the quality of student prose. A number of institutions and professional associations have suggested or already

implemented in-service workshops to train faculty in evaluating and teaching composition skills. Such developmental efforts will certainly be beneficial; but we do not believe that they are a necessary prelude to a general increase in the amount of writing to be done by students and the amount of feedback on that writing to be provided by faculty.

One of the graders in the study wrote a general comment on the score sheet. "While these papers vary in quality (both with regard to content and language) none seems to come close to the 'awful stuff' which I sometimes get, especially at the freshman and sophomore levels. Does this say anything about the sample?" We do not believe that there was anything wrong with the sample. Rather, the same students who write reasonably competent prose on a composition proficiency examination will write awful stuff for a faculty member who apparently is willing to accept it. We believe that students and faculty are both capable of rising to higher levels of expectancy than have been previously held out to them.